Real Estate Investment in Germany

IBM Data Science Capstone Project

Xia Ma, March 05, 2021

Introduction: Business Problem

In this project, I want to find out the best city for investing in real estate market in Germany. Specifically, this report will be targeted to the stakeholders who are interested in **Residential Rental Properties** and planing to make a profitable investment in Germany.

Since there are too many cities, we need to narrow our research down to large cities, which, defined by *German Federal Institute for Research on Building, Urban Affairs and Spatial Development*, should have more than 100,000 inhabitants. We assume that large cities in the long term will generate higher ROI (return of investment) than small cities or even villages.

There are a lot of important factors for a successful real estate investment, such as location, population and economy, purchase price and rental yields etc. In this project, we won't consider some common factors, such as maintenance, insurance, tax and loan, since those costs are independent of where the property is located and therefore won't influence our decision.

We will use our data science skills to evaluate different cities based on certain criteria. Advantages and disadvantages of each city will then be clearly expressed so that possible final decision could be made by stakeholders.

Data Description

Based on the definition of our problem, important factors that will influence our investment decision in a certain city are: **Economic Strength, Population, Rental Yields Estimation and Average Purchase Price**.

To determine the city candidates, I used the data collected from the following four sources:

- First, a property can better hold its value from a long-run perspective, if it's located in a region, which has a strong economy and enjoys high standards of living. Wikipedia just provides a great list of all German cities by GDP (gross domestic product). So, I will use this to determine the economic strength of each city. [1]
- Secondly, a large population and continuous population growth trend will lead to large housing demand, from which the investment will benefit a lot. Therefore, it's critical that the city candidates should keep attracting new inhabitants rather than losing some. I found a list of German cities by population from Wikipedia. [2]

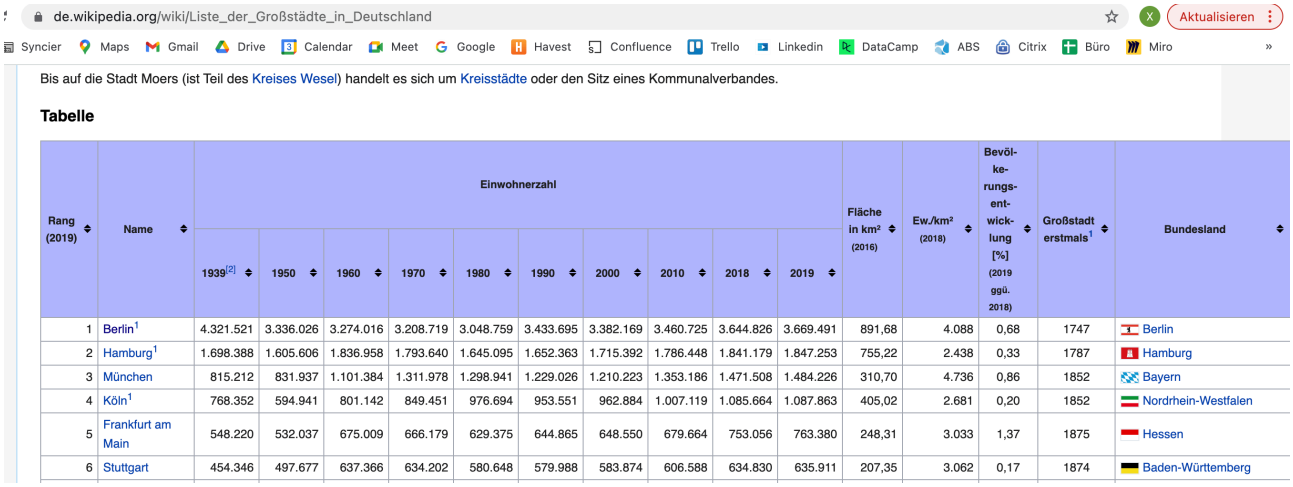
To calculate ROI (return of investment), we need to know the return on the one hand, and investment on the other hand.

- For all residential rental properties in Germany, the rent index is one of the legally provided options for determining the local comparative rent in privately financed residential construction. It serves as a justification for rent increases and is drawn up by cities. If interested, you can look for details on their official websites. In this project, we will use this index to determine the possible rental yields. For simplicity, I will directly use the data on the website of Wohnung. [3]
- Purchase price depends on many factors, for example whether it's a house or an apartment, a new construction or existing property. Still for simplicity, I will only focus on existing apartments with 3 rooms and ca. 80 m^2, which are listed on the website of DasHaus. [4]

Data Preprocessing

In this section, I will collect and preprocess the data from the sources mentioned above. However, I wanted to challenge myself this time. Instead of doing copy-paste into a file in the traditional way, I wanted to scrape the data directly from HTML tables into a data frame using Python web scraping skills. For this, there are different libraries and packages in Python. For the first two data sources, I applied the more complicated method with BeautifulSoup package. For the third and fourth sources, I used requests function and transform the data into data frame using pandas. The results are four nice data frames.

There are usually a lot of useless information like hyperlinks or pictures on websites. I preselected the necessary data with BeautifulSoup, in order to minimize efforts of data cleaning at a later stage. The Wikipedia page looks like this:



Rang (2019)	Name	Einwohnerzahl										Fläche in km² (2016)	Ew./km² (2018)	Bevölkerungsentwicklung [%] (2019 ggü. 2018)	Großstadt erstmals[1]	Bundesland
		1939[2]	1950	1960	1970	1980	1990	2000	2010	2018	2019					
1	Berlin[1]	4.321.521	3.336.026	3.274.016	3.208.719	3.048.759	3.433.695	3.382.169	3.460.725	3.644.826	3.669.491	891,68	4.088	0,68	1747	Berlin
2	Hamburg[1]	1.698.388	1.605.606	1.836.958	1.793.640	1.645.095	1.652.363	1.715.392	1.786.448	1.841.179	1.847.253	755,22	2.438	0,33	1787	Hamburg
3	München	815.212	831.937	1.101.384	1.311.978	1.298.941	1.229.026	1.210.223	1.353.186	1.471.508	1.484.226	310,70	4.736	0,86	1852	Bayern
4	Köln[1]	768.352	594.941	801.142	849.451	976.694	953.551	962.884	1.007.119	1.085.664	1.087.863	405,02	2.681	0,20	1852	Nordrhein-Westfalen
5	Frankfurt am Main	548.220	532.037	675.009	666.179	629.375	644.865	648.550	679.664	753.056	763.380	248,31	3.033	1,37	1875	Hessen
6	Stuttgart	454.346	497.677	637.366	634.202	580.648	579.988	583.874	606.588	634.830	635.911	207,35	3.062	0,17	1874	Baden-Württemberg

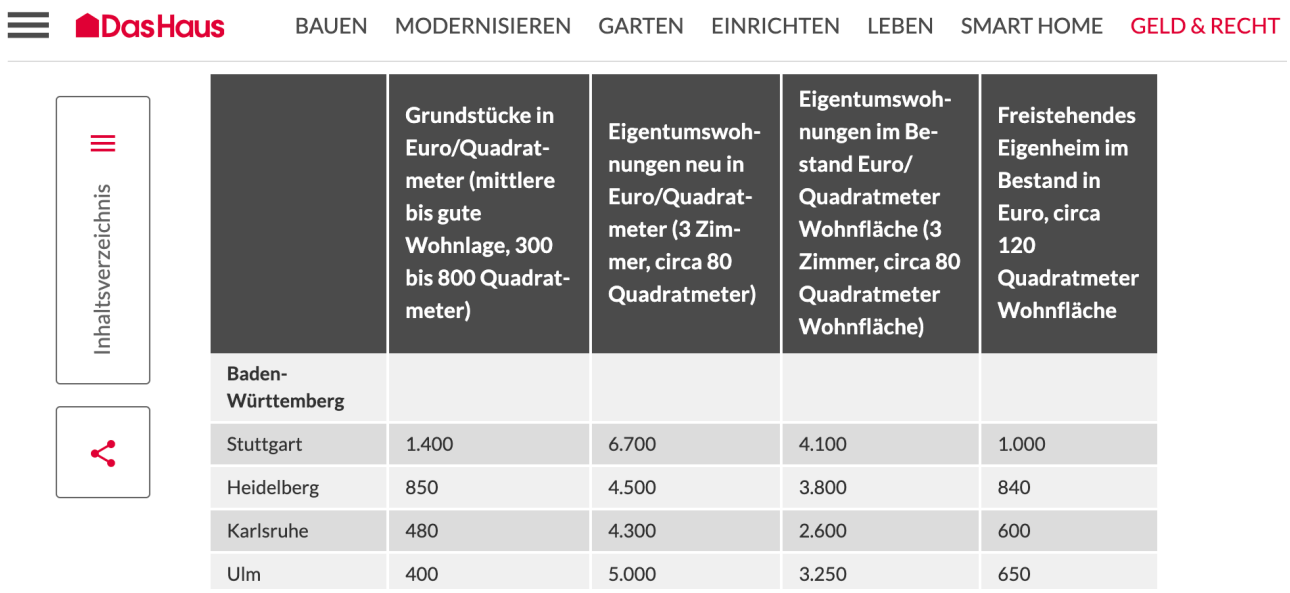
After analyzing the source code of the page in HTML, I recognized how the code structure looks like and where the desired table is located.

In order to get clean data for later analysis, I made some effort to convert data types, select columns, reset index and adjust texts etc. The result is a list of cities with their population in 2000 and 2019 as well as land area. With these information, I can calculate

the population density and population growth in the last two decades. After that, I went on extracting GDP data from Wikipedia in the same way.

City	Census_2000	Census_2019	Land_Area_in_km²
Berlin	3382169	3669491	891.68
Hamburg	1715392	1847253	755.22
München	1210223	1484226	310.70
Köln	962884	1087863	405.02
Frankfurt am Main	648550	763380	248.31
Stuttgart	583874	635911	207.35

For the data about property purchase price and rental yields, I simply used Pandas to read the whole table from the websites into data frames and do data cleaning later using pandas:



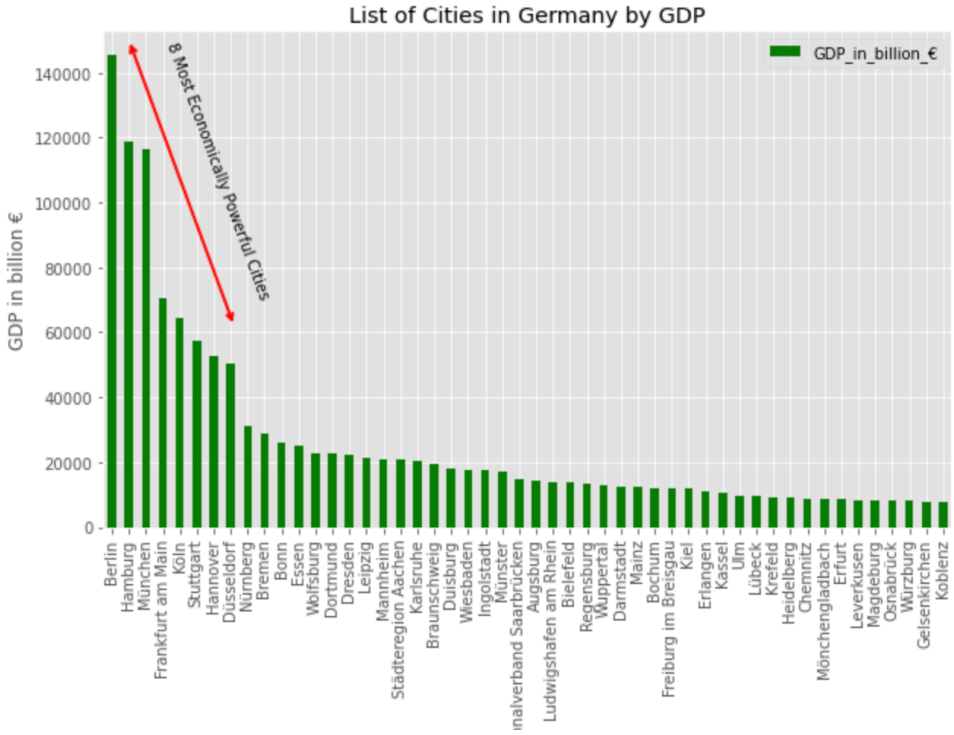
	Grundstücke in Euro/Quadrat-meter (mittlere bis gute Wohnlage, 300 bis 800 Quadrat-meter)	Eigentumswoh-nungen neu in Euro/Quadrat-meter (3 Zim-mer, circa 80 Quadratmeter)	Eigentumswoh-nungen im Be-stand Euro/ Quadratmeter Wohnfläche (3 Zimmer, circa 80 Quadratmeter Wohnfläche)	Freistehendes Eigenheim im Bestand in Euro, circa 120 Quadratmeter Wohnfläche
Baden-Württemberg				
Stuttgart	1.400	6.700	4.100	1.000
Heidelberg	850	4.500	3.800	840
Karlsruhe	480	4.300	2.600	600
Ulm	400	5.000	3.250	650

	Unnamed: 0	Grundstücke in Euro/Quadrat- meter (mittlere bis gute Wohnlage, 300 bis 800 Quadrat- meter)	Eigentumswoh- nungen neu in Euro/Quadrat- meter (3 Zim- mer, circa 80 Quadratmeter)	Eigentumswoh- nungen im Be- stand Euro/ Quadratmeter Wohnfläche (3 Zimmer, circa 80 Quadratmeter Wohnfläche)	Freistehendes Eigenheim im Bestand in Euro, circa 120 Quadratmeter Wohnfläche
0	Baden-Württemberg	NaN	NaN	NaN	NaN
1	Stuttgart	1400.0	6700.0	4100.0	1000.0
2	Heidelberg	850.0	4500.0	3800.0	840.0
3	Karlsruhe	480.0	4300.0	2600.0	600.0
4	Ulm	400.0	5000.0	3250.0	650.0
5	Bayern	NaN	NaN	NaN	NaN
6	München	2300.0	8100.0	6900.0	1500.0

Methodology

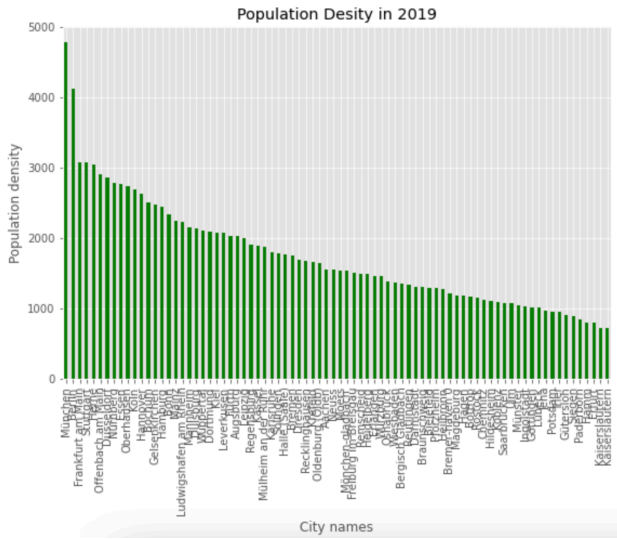
As mentioned above, there are four important factors for our investment decision: economic power, population, rental yields and average purchase price. All the data are now stored in corresponding data frames: gdp_data, population_data, purchase_price and rental_yield. In this part, I will describe the data analysis step by step and show how I used the data to get the final result.

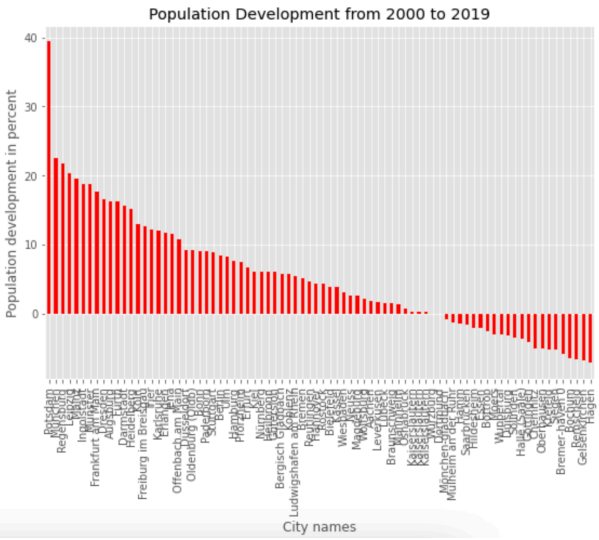
First of all, we need to narrow down city candidates based on GDP and population data. GDP ranking list is already prepared. It's pretty clear, which cities are economically stronger. The top-8 cities are striking and marked with a red line:



Concerning the population data, we need to consider both the current population and the growth trend, since there are many large cities whose population however shrink a lot in the last few decades, for example those cities in Ruhr (a traditional heavy industry region in West Germany).

In the following chart on the right, we can see that about 50 metropolitan cities keep attracting new people in the last 20 years, while the rest of them are all losing people. Therefore, this overview helps us set proper conditions to kick those shrinking cities out of our research scope and only focus on cities whose population is not only large, but also keeps growing very fast.

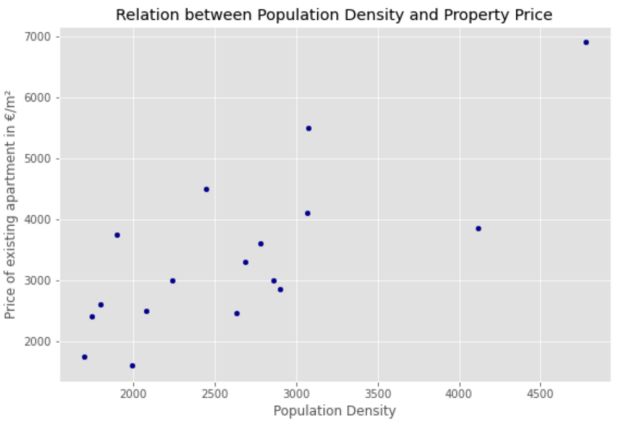


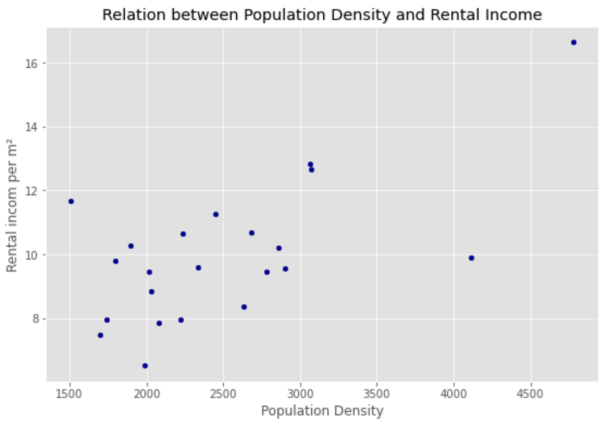


We've already got the purchase prices and rental yields from the data preprocessing step. After dividing return by investment, we got ROI. In the following data frame, I consolidated all the data including GDP, population, property price, rental income and calculated ROI.

City	GDP_in_billion_€	Population_Density	Population_Development_in_Percent	Price_of_existing_apartment_in_€/m²	Rental_income_per_m²	ROI_in_%
Berlin	145547.0	4115.3	8.5	3850.0	9.89	2.5
Hamburg	118912.0	2446.0	7.7	4500.0	11.25	2.5
München	116647.0	4777.0	22.6	6900.0	16.64	2.4
Frankfurt am Main	70639.0	3074.3	17.7	5500.0	12.66	2.3
Köln	64536.0	2685.9	13.0	3300.0	10.68	3.2

With all needed information in one data frame, I was able to explore the data a little and tried to find some connection between different factors. In the following scatter plots, we can for example see a positive correlation between population density and property price as well as between population density and rental income. This makes sense. The more people there are, the bigger the housing demand will be.



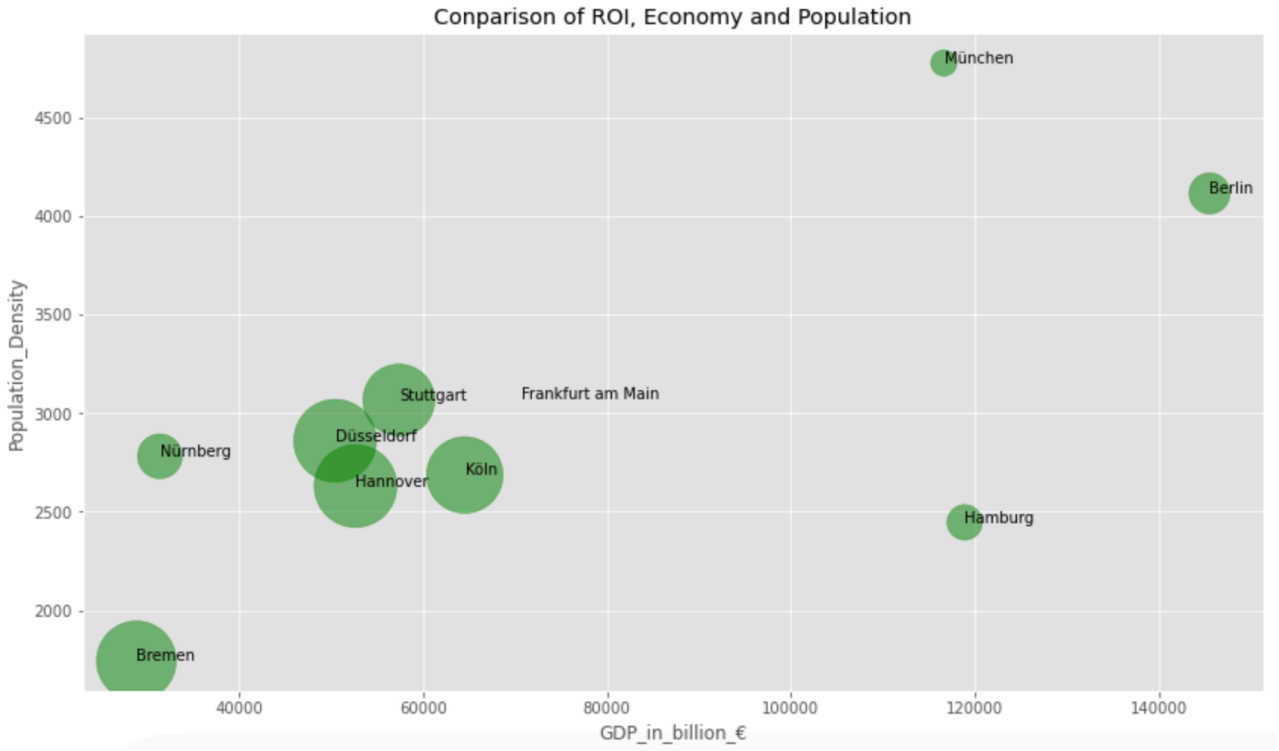


Results and Discussion

After all those steps of data collection, preprocessing and exploratory analysis, we are finally there to show the result. In the following table, all the 9 candidates are listed, which fulfill all of our criteria: economically the strongest metropolises in Germany with the largest population and fastest population growth trend.

City	GDP_in_billion_€	Population_Density	Population_Development_in_Percent	Price_of_existing_apartment_in_€/m²	Rental_income_per_m²	ROI_in_‰
Berlin	145547	4115.3	8.5	3850	9.89	2.57
Hamburg	118912	2446.0	7.7	4500	11.25	2.50
München	116647	4777.0	22.6	6900	16.64	2.41
Frankfurt am Main	70639	3074.3	17.7	5500	12.66	2.30
Köln	64536	2685.9	13.0	3300	10.68	3.24
Stuttgart	57369	3066.8	8.9	4100	12.84	3.13
Hannover	52655	2630.2	4.3	2460	8.37	3.40
Düsseldorf	50429	2860.4	9.2	3000	10.20	3.40
Nürnberg	31374	2781.3	6.1	3600	9.44	2.62
Bremen	28818	1743.3	5.2	2400	7.96	3.32

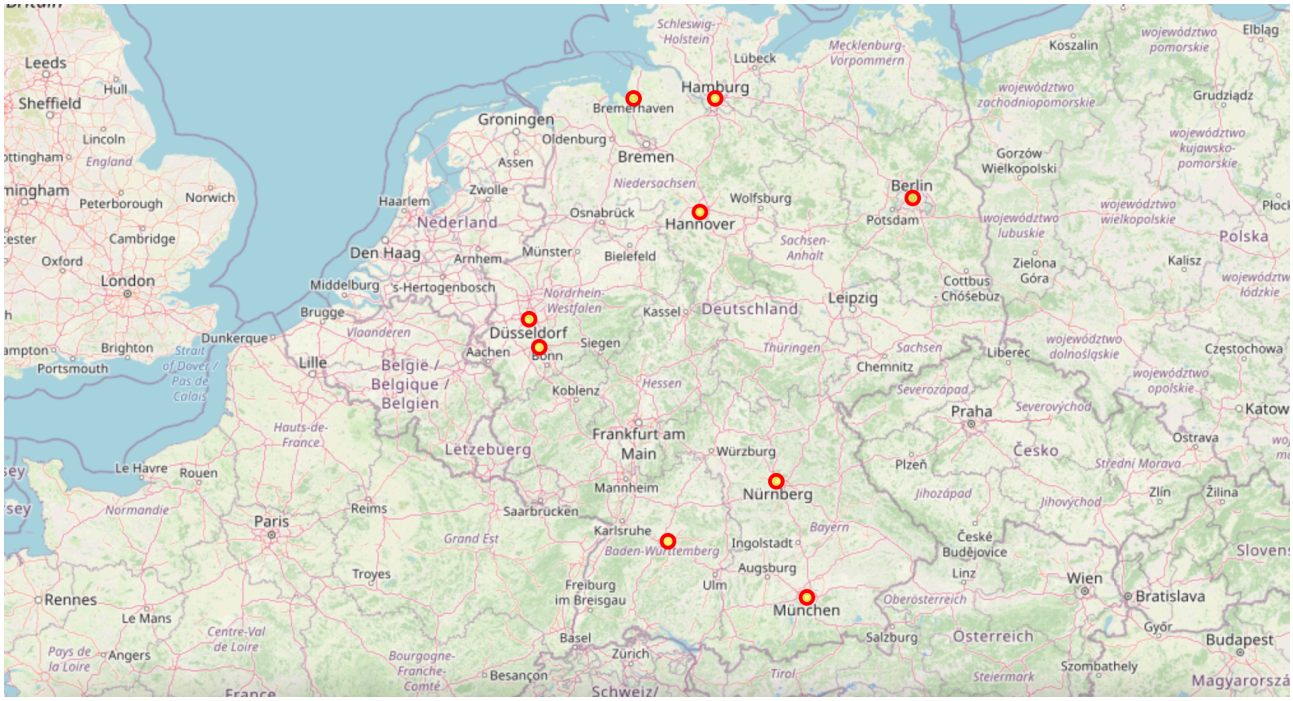
In order to better recognize the differences among candidates and choose the best one for our investment, I demonstrated them in a bubble chart, with GDP on the x-axis, population density on the y-axis and the size of the bubbles representing the ROI.



It's quite interesting to visualize the result in graphs. While the megacities like Munich, Berlin and Hamburg don't show a better ROI, the second-largest cities such as Hannover, Dusseldorf, Stuttgart and Cologne seem to be more promising.

One possible explanation for this could be: although these super metropolises have much larger population and stronger economy, they are also quite crowded, which leads to very high property price and thus reduces the ROI.

Finally, I used Folium to demonstrate these cities on the map of Germany, so that you can see where they are all located. Maybe this will help with your investment decision :)



Conclusion

For me, this capstone project of IBM is more of a self-assessment-challenge than serves a real purpose. I learned a lot from this course and tried out different tools and skills.

At the end of this project, every single step of finishing this assignment comes to my mind: from data scraping, cleaning and preprocessing, transforming and analyzing to visualizing. How many times I ran into difficulties. How many times I then found useful advices from the internet and even borrowed some fragments of codes and adjusted them to my needs. It's really hard work here, but it's definitely worth it.

All the tools are simply there. You just need to know how to use them at work or in everyday life. Even though you may sometimes have trouble, there are so many excellent generous people in the internet, who can help.

My notebook with all technical codes for this project is also available to the public on Github [5]. Feel free to contact me if you have any questions or comments. Looking forward to talking to you about data science!

References

[1] https://de.wikipedia.org/wiki/Liste_der_deutschen_St%C3%A4dte_nach_Bruttoinlandsprodukt

[2] https://de.wikipedia.org/wiki/Liste_der_Gro%C3%9Fst%C3%A4dte_in_Deutschland

[3] https://www.wohnung.com/mietpreise

[4] https://www.haus.de/geld-recht/immobilienpreise-deutschland-vergleich-2019

[5] https://github.com/xia-ma/coursera-capstone/blob/master/Capstone_Final_Project.ipynb